



Bacterial strain Taxonomy for Genomic Surveillance

Enterobase and Hierarchical Clustering

Nigel Dyer: University of Warwick, UK

Created: October 2025

Outline

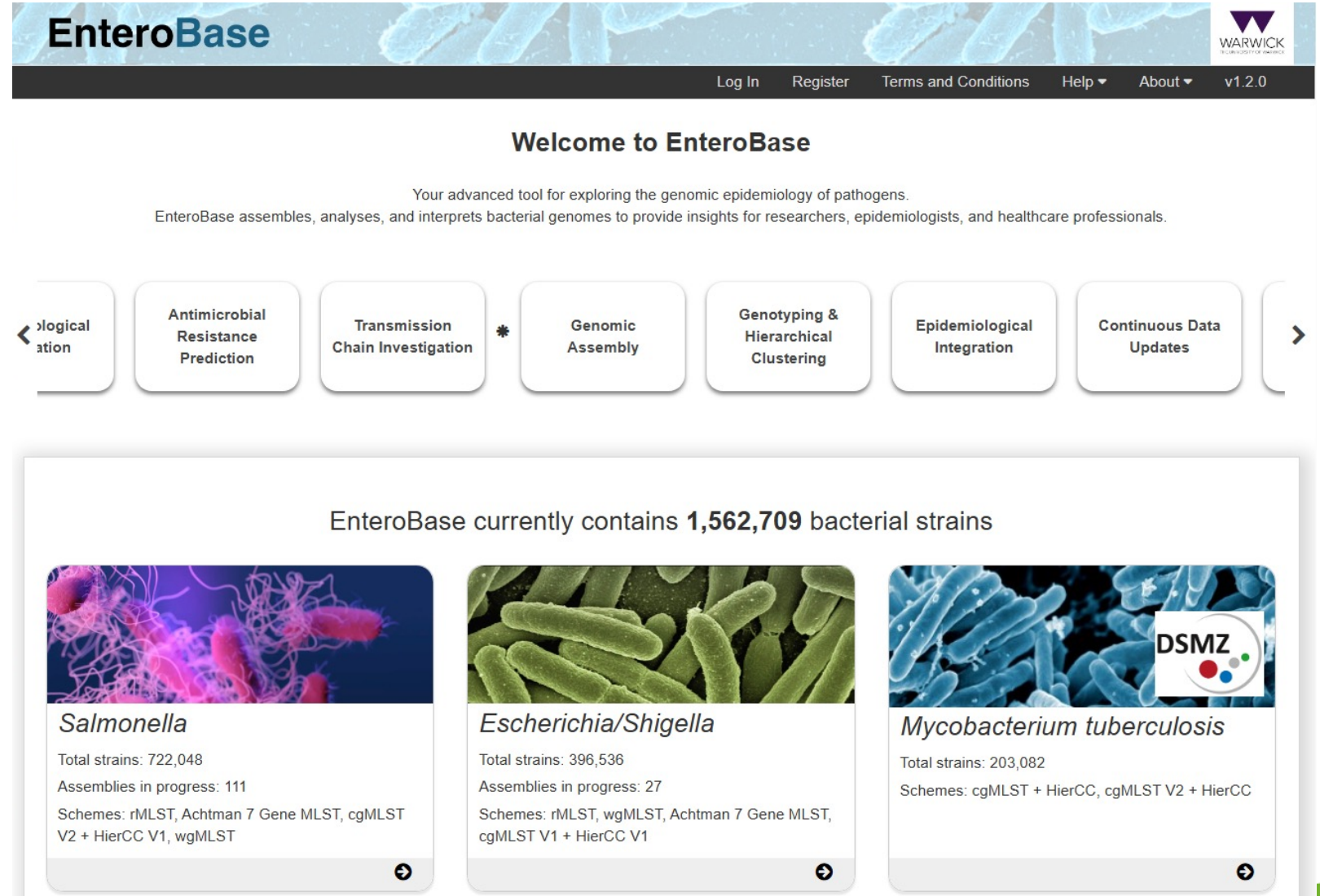
This session consists of the following elements

1. Introduction to Enterobase
2. Hierarchical clustering and LIN codes
3. Usage examples

EnteroBase — <https://enterobase.warwick.ac.uk/>

Species:

- Salmonella
- Escherichia
- Streptococcus
- Clostridioides
- Vibrio
- Helicobacter
- Yersinia
- Moraxella



The screenshot shows the EnteroBase website interface. At the top is a navigation bar with the EnteroBase logo, a Warwick University logo, and links for Log In, Register, Terms and Conditions, Help, About, and version v1.2.0. Below the navigation bar is a welcome message: "Welcome to EnteroBase" followed by a description: "Your advanced tool for exploring the genomic epidemiology of pathogens. EnteroBase assembles, analyses, and interprets bacterial genomes to provide insights for researchers, epidemiologists, and healthcare professionals." A horizontal menu of tool categories is displayed: Logical annotation, Antimicrobial Resistance Prediction, Transmission Chain Investigation, Genomic Assembly (highlighted with an asterisk), Genotyping & Hierarchical Clustering, Epidemiological Integration, and Continuous Data Updates. Below this menu, a section states "EnteroBase currently contains 1,562,709 bacterial strains". Three featured bacterial species are shown in cards: Salmonella (722,048 strains), Escherichia/Shigella (396,536 strains), and Mycobacterium tuberculosis (203,082 strains). Each card includes a microscopic image, the species name, total strains, assemblies in progress, and the schemes used for analysis.

EnteroBase

Log In Register Terms and Conditions Help About v1.2.0

Welcome to EnteroBase

Your advanced tool for exploring the genomic epidemiology of pathogens.
EnteroBase assembles, analyses, and interprets bacterial genomes to provide insights for researchers, epidemiologists, and healthcare professionals.

Logical annotation Antimicrobial Resistance Prediction Transmission Chain Investigation * Genomic Assembly Genotyping & Hierarchical Clustering Epidemiological Integration Continuous Data Updates

EnteroBase currently contains **1,562,709** bacterial strains



Salmonella

Total strains: 722,048
Assemblies in progress: 111
Schemes: rMLST, Achtman 7 Gene MLST, cgMLST V2 + HierCC V1, wgMLST



Escherichia/Shigella

Total strains: 396,536
Assemblies in progress: 27
Schemes: rMLST, wgMLST, Achtman 7 Gene MLST, cgMLST V1 + HierCC V1



Mycobacterium tuberculosis

Total strains: 203,082
Schemes: cgMLST + HierCC, cgMLST V2 + HierCC



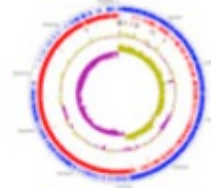
Are you familiar with EnteroBase?



User reads



Sequence read archive



Complete genomes



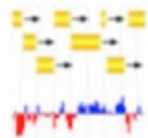
Metadata



EnteroBase

Curated database
Web based GUI

Standardized toolkit
Analysis workspaces
API



Genome assembly
& annotation



Sequence typing,
strain classification



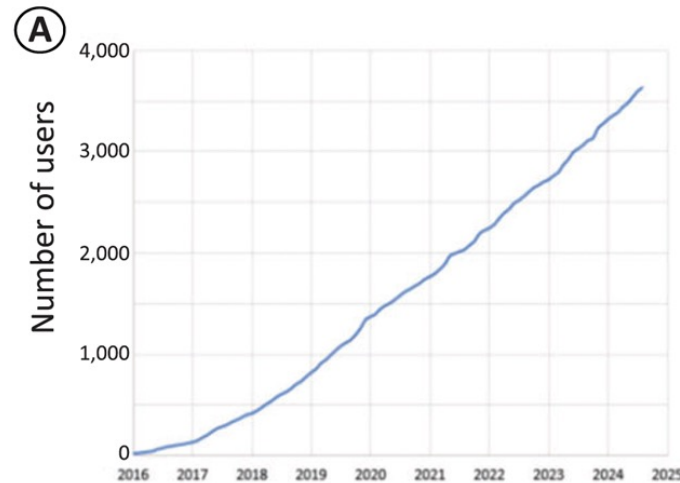
Phylogenetic analysis



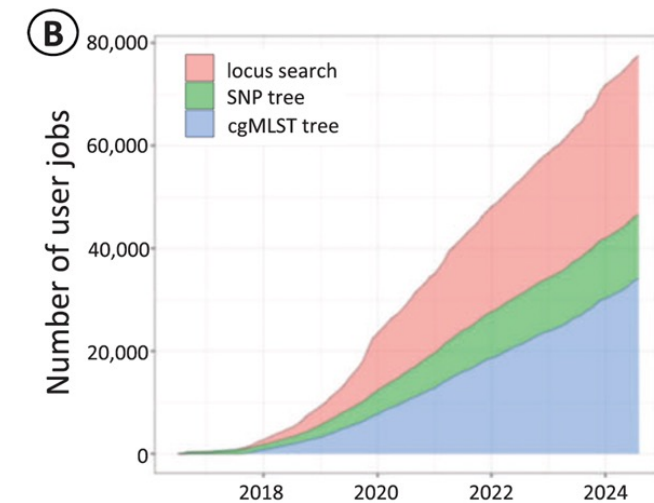
AMR genotyping

Usage statistics

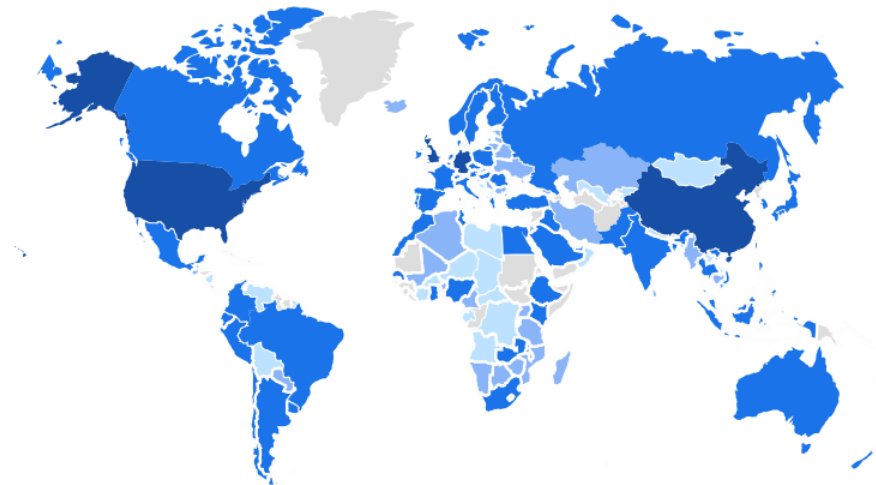
(A) Registered EnteroBase users, 2016-2024



(B) Cumulative number of user jobs



(C) User access statistics collected by Google Sep 24 to Sep 25. Table shows top seven countries



United Kingdom	6.3K
China	3.4K
United States	3.1K
Germany	2.2K
France	1K
Japan	847
India	654

~4000 users worldwide

Including Government Agencies:

- US Department of Agriculture
- Public Health England
- Institut Pasteur (France)
- Animal Health and Veterinary Laboratories Agency (UK)
- Antibiotic Resistance Monitoring and Reference Laboratory (UK)
- National Salmonella Reference Library (Ireland)
- Bavarian Health and Food Safety Authority
- Laboratoire National de Sante (Luxembourg)
- National Reference Laboratories (Israel)
- Public Health Agency of Sweden
- Norwegian National Advisory Unit on Detection of Antimicrobial Resistance
- National Institute of Health (Portugal)
- Centre for Zoonoses and Environmental Microbiology (Netherlands)
- Australian National Salmonella Reference Laboratory
- Centre for Infectious Diseases & Microbiology (Sydney, Australia)
- CDC (Taiwan), etc etc etc

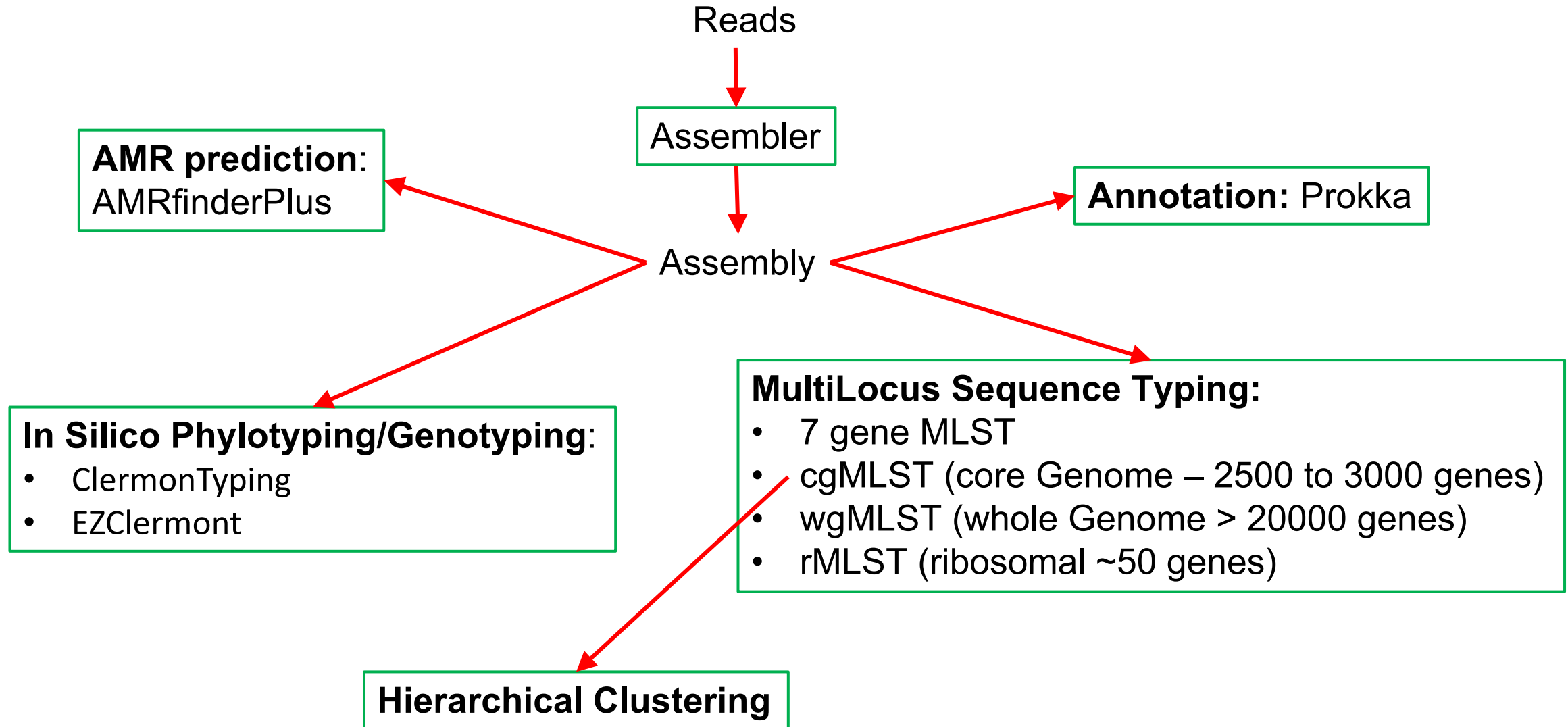
UNIVERSITY
OF WARWICK



Public Health
England



Enterobase data flow and analysis



EToKi Assembler (EnteroBase Tool Kit)

- <https://github.com/zheminzhou/EToKi>
- SPAdes-based assembly pipeline, with improvements:
- Pre-processing “prepare” step:
 - Trims sequences based on base-qualities (bbduk)
 - Removes potential adapters and barcodes (bbduk)
 - Limits total amount of reads to be used to 120X
- “Assemble” step:
 - Assembles short reads (SPAdes)
 - Maps reads back to assembled genome (Minimap2)
 - Polishes consensus using hapog
 - Removes low level contaminations (contigs with <30% of average read coverage)
 - Estimates the base quality of the consensus
 - Predicts taxonomy using Kraken (MiniKraken)

Assembly tool evaluation exercise for UK government project PATHSAFE

(Pathogen Surveillance in Agriculture, Food and the Environment)

EToKi outperformed:

- SPAdes (default settings)
- MEGAHIT
- IDBA
- Unicycler
- Velvet
- SKESA
- Masurca
- PathogenWatch

Multiple assessment criteria including comparison of post-assembly in silico MLST genotyping

Comparison of hierCC and cgMLST LIN codes

- Strains allocated to clusters using identical algorithm
- Same algorithm for adding strains to existing clusters
- Same algorithm for creating new clusters
- Different ways of identifying the same clusters
- No cluster fusion
- Cluster definitions are stable

Comparison of hierCC and cgMLST LIN codes

- 1:1 mapping between cluster identifiers.

LIN codes	hierCC
3-0-1-2-3, 3-0-1-2-4	34,56
Explicit cluster hierarchy	Cluster hierarchy not obvious in cluster identifiers
Cluster at level N identified by unique sequence of N numbers	Cluster at any level identified by a unique number. Numbers reused between levels.

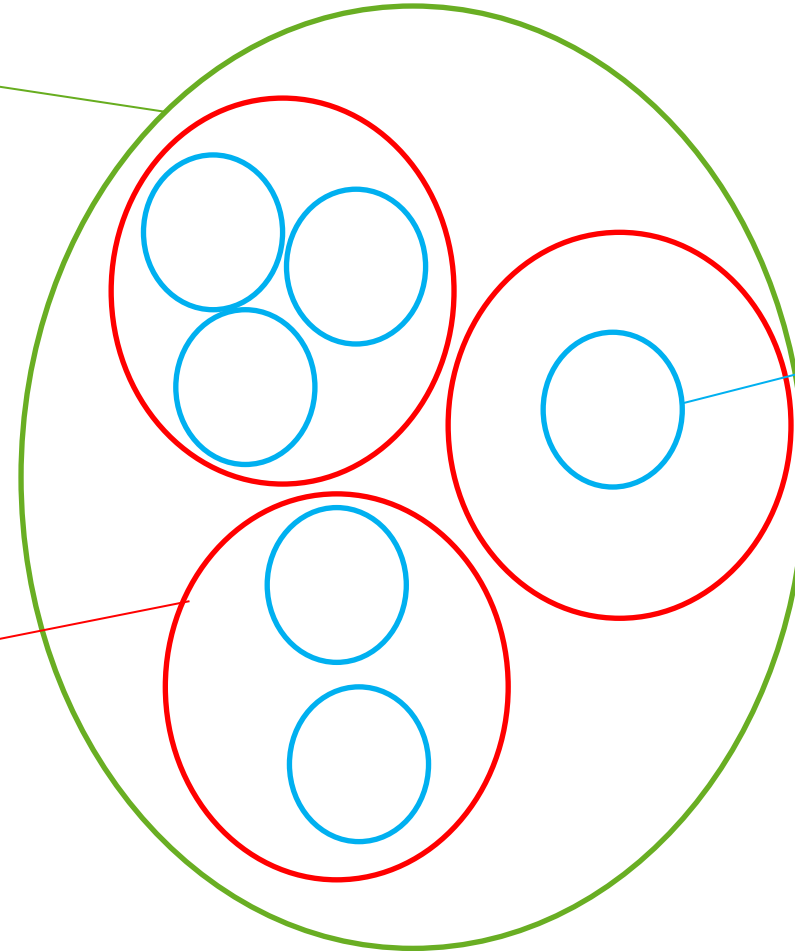
Comparison of hierCC and LIN codes

hc200:

Strains differ
by fewer than
200 alleles

hc100:

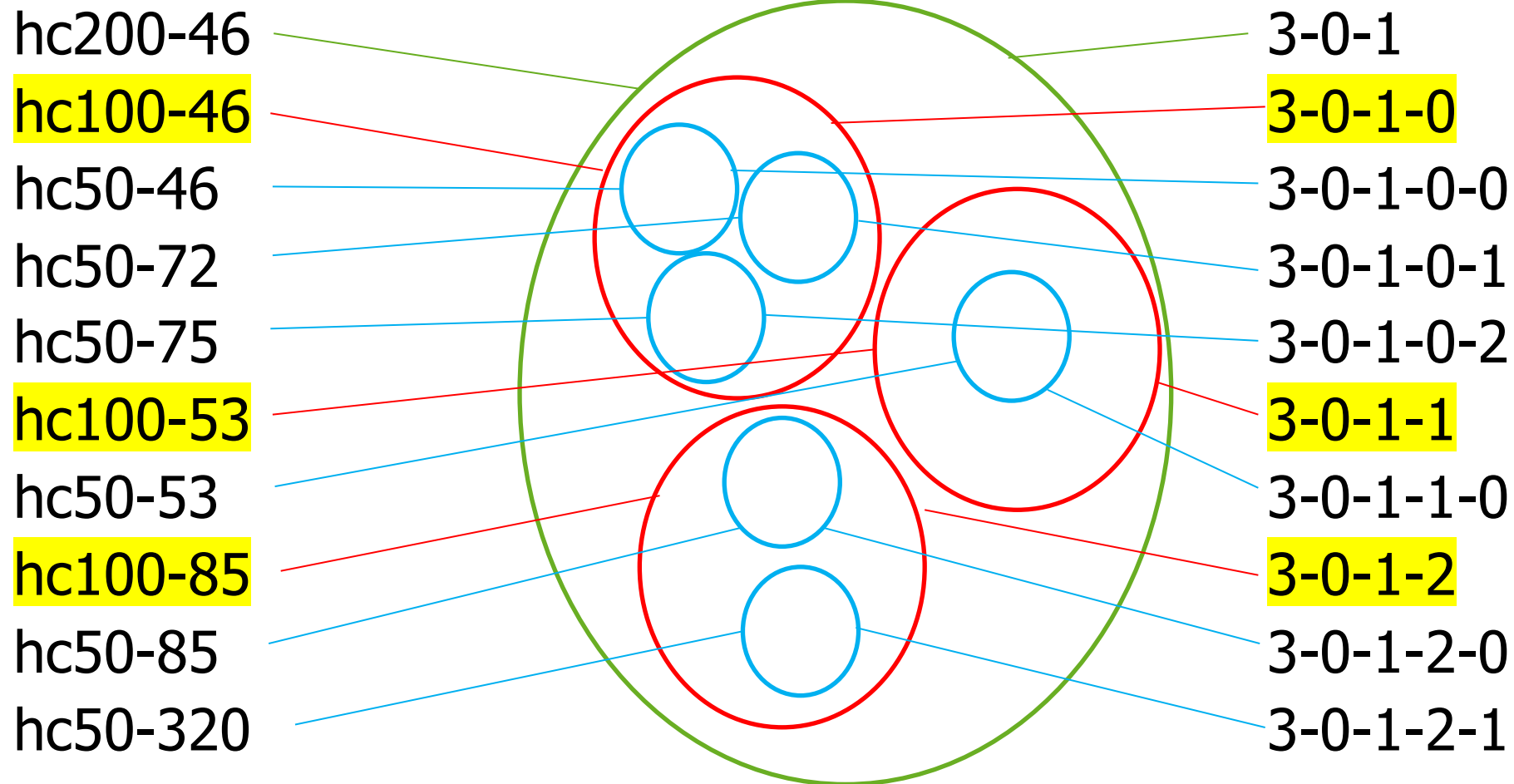
Strains differ
by fewer than
100 alleles



hc50:

Strains differ by
fewer than 50
alleles

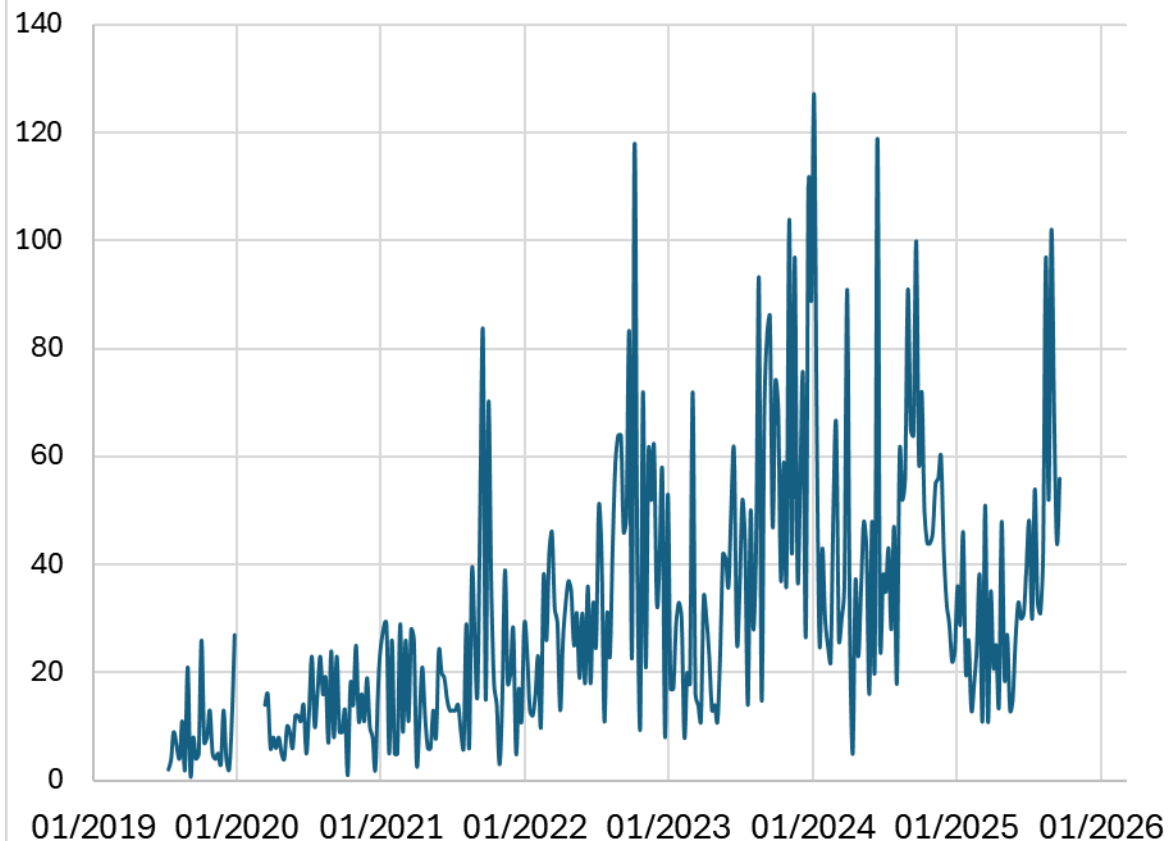
Comparison of hierCC and LIN codes



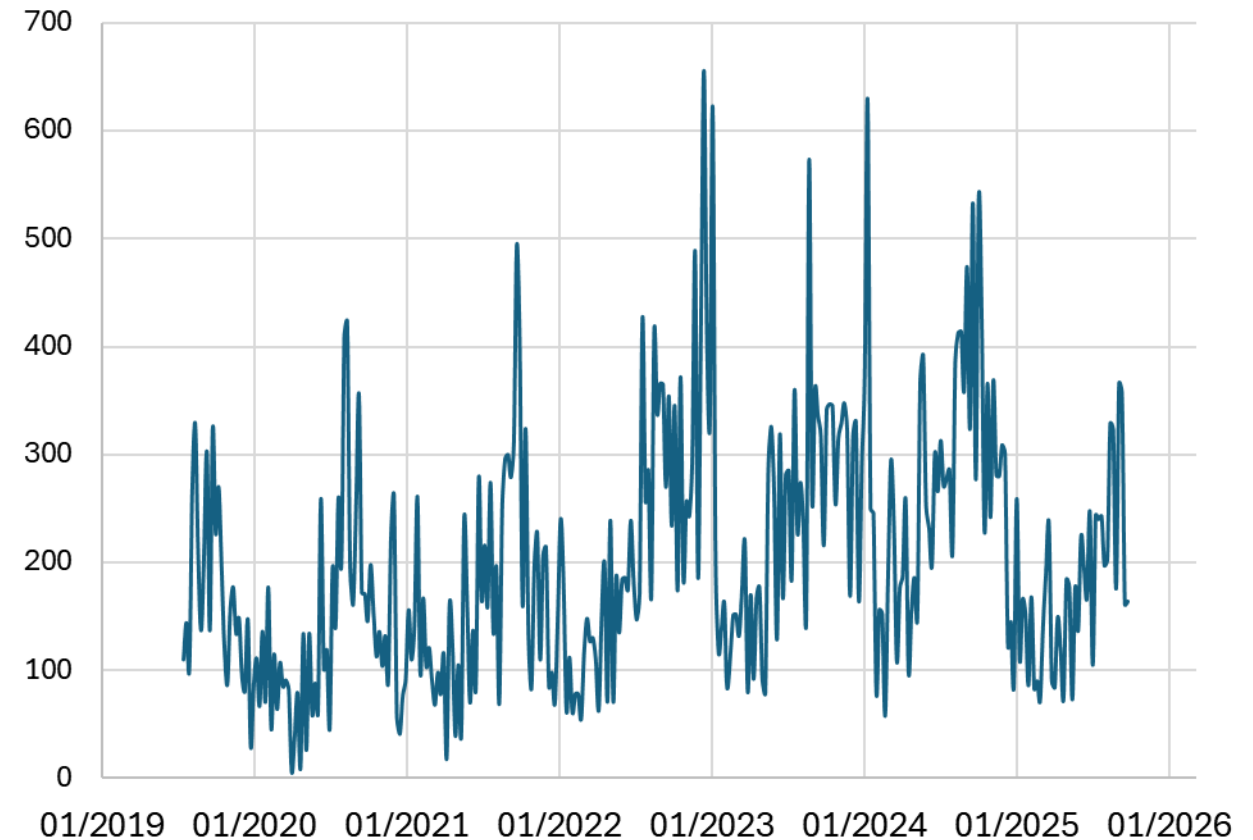
Outbreak surveillance: Institut Pasteur

Weekly strain submission

E. Coli/Shigella



Salmonella



> 10 000 clinical isolates per year (hospitals and private labs)



Isolation in **XLD** (Xylose Lysine Deoxycholate Agar)

- **Hajna & Chromagar** (in case of contaminations and/or atypical colonies)
- **MALDI-TOF MS** (genus confirmation in case of atypical colonies)
- **Seroagglutination** (in case of several isolates for the same patient)



Whole-genome sequencing (Illumina NextSeq500) & molecular typing:

- Species identification
- MLST (7 locus-scheme, Achtman et al. *PLoS Pathog* 2012)
- resistome/virolome (ResFinder, PlasmidFinder...)
- *in silico* agglutination patterns (***Salmotyper***: Fabre, in prep.)
- **cgMLST** (3,002 genes) - (Alikhan et al. *PLoS Genet* 2018)



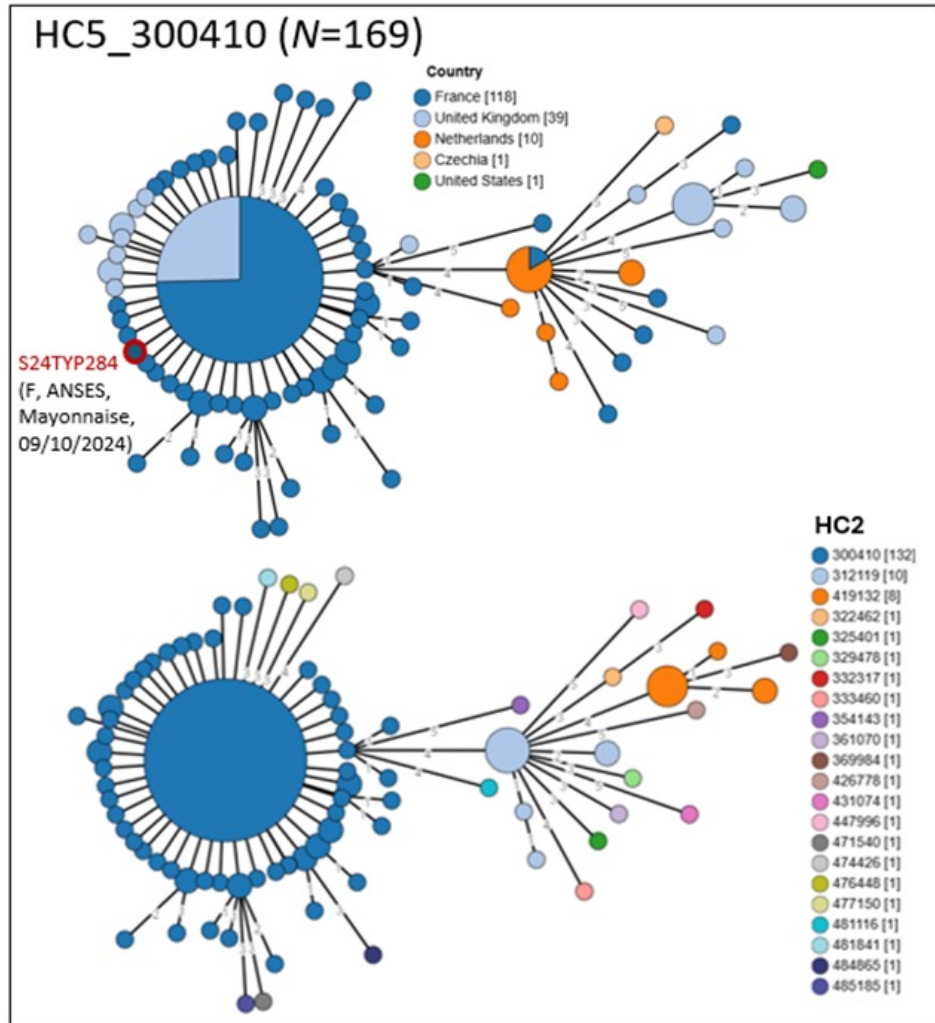
Identification of clusters cgMLST (HC2/HC5 with >8 cas in a 8 weeks timeframe)

Communication of clusters to **Santé publique France** for investigation in collaboration with food authorities (DGAL, MUS, etc.).



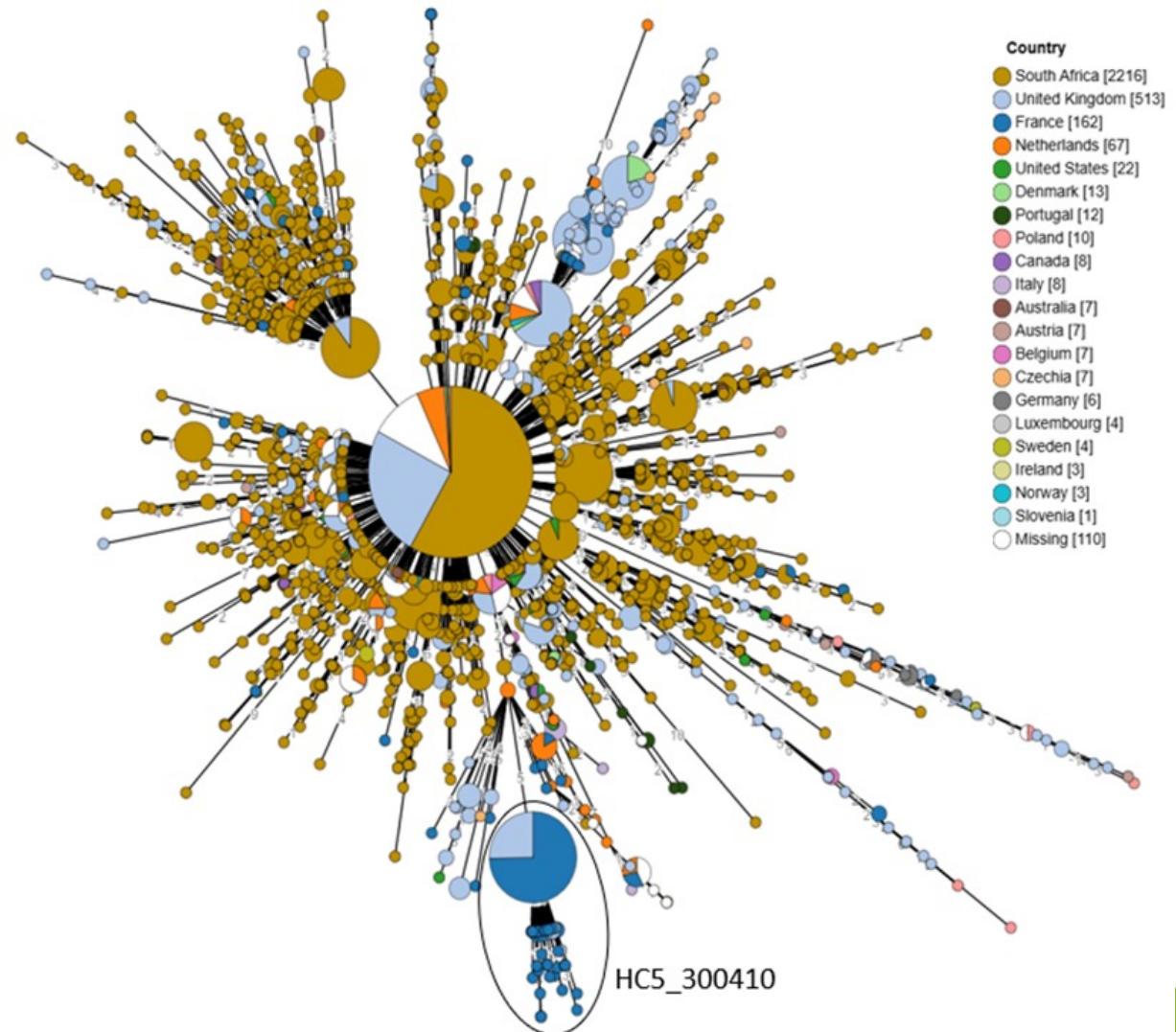
- **wgSNP analysis** (Enterobase; snippy) – confirmation of cgMLST clusters at whole genome level (publications, court investigations, etc.)
- **Long-reads sequencing** (ONT Nanopore) – reconstruction chromosomes and plasmides

Oct-Nov 2024 Salmonella outbreak



cgMLST Enterobase (3,002 loci)

HC10_2301 (N=3,190)



One day's data from France: hierCC identifiers

Enterobase

NigelDyer Terms and Conditions Help About v1.2.0

Salmonella

Admin Tasks

Job Monitor

Admin Page

Email Users

Tasks

Search Strains

Find ST(s)

Download Schemes

Upload Reads

Curate Strains

My Jobs

Locus Search

Accessory Genome

My Buddies

Search Strains ¹

Data View Workspace Experiment Workspace:None Rows Total:123 Filtered:123 Maximize

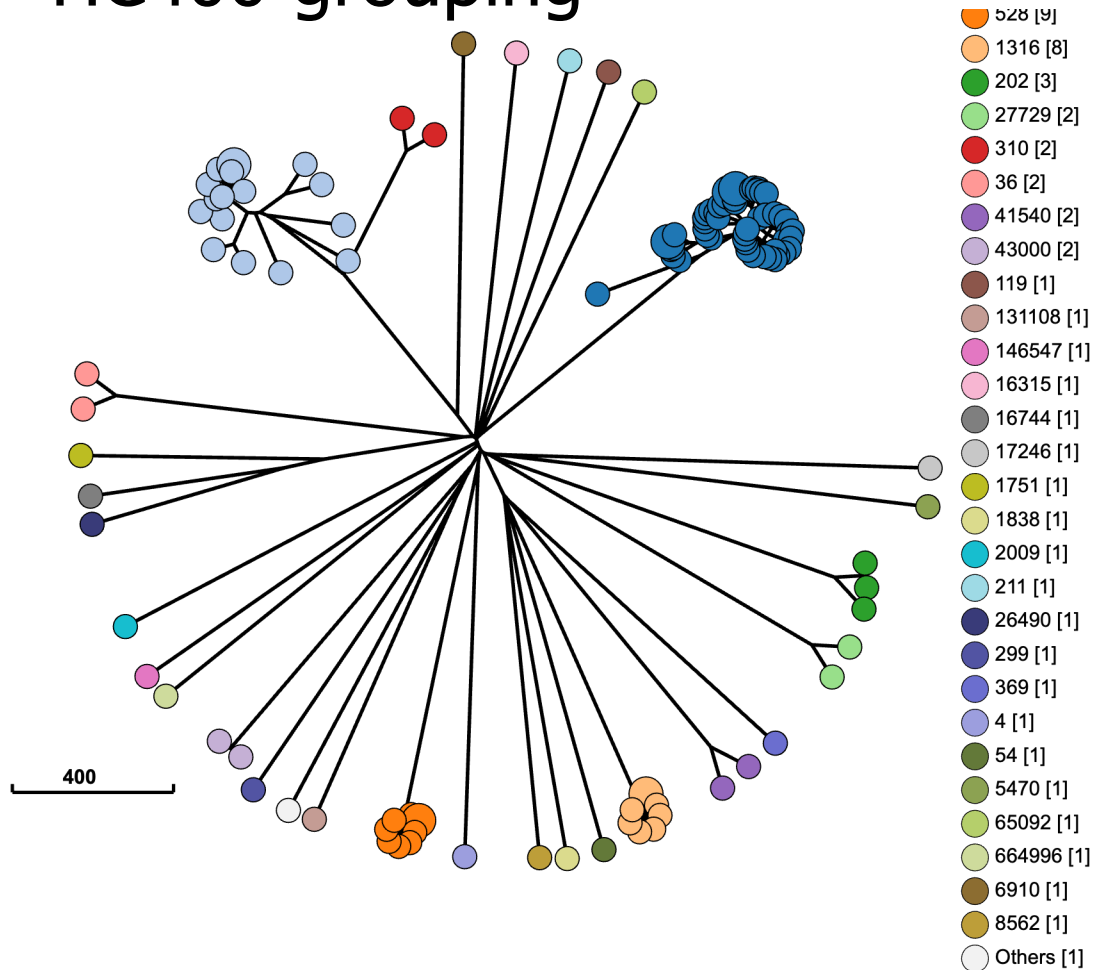
Edit Mode: ☐ Experimental Data cgMLST V2 + HierCC V1

Uberstrain	✓	✎	Name	ST	HC0 (in...	HC2	HC5	HC10	HC20	HC50	HC100	HC200	HC400	HC
■SAL_CD5...			202507889	663977	663977	663977	1664	528	528	528	528	528	528	528
■SAL_CD5...			202507891	650032	650032	650032	650032	366935	338501	338501	5470	5470	5470	5470
■SAL_CD5...			202507899	663985	663985	663985	147160	147160	56084	27753	27753	12	12	12
■SAL_CD5...			202507892	663972	663972	2301	2301	1398	397	397	12	12	12	12
■SAL_CD5...			202507907	663975	663975	295559	286848	1316	1316	1316	1316	1316	1316	1316
■SAL_CD5...			202507894	663980	663980	663980	385183	385183	385183	189533	41540	41540	41540	41540
■SAL_CD5...			202507883	663986	663986	469974	448749	202	202	202	202	202	202	202
■SAL_CD5...			202507890	663974	663974	238926	2301	1398	397	397	12	12	12	12
■SAL_CD5...			202507896	663987	663987	663987	124098	124098	7820	1404	1404	1404	12	12
■SAL_CD5...			202507888	663973	663973	363896	248308	1398	397	397	12	12	12	12
■SAL_CD5...			202507880	663976	663976	469161	1316	1316	1316	1316	1316	1316	1316	1316
■SAL_CD5...			202507882	663982	663982	469974	448749	202	202	202	202	202	202	202
■SAL_CD5...			202507902	664023	664023	663985	147160	147160	56084	27753	27753	12	12	12
■SAL_CD5...			202507906	664004	664004	468804	253286	56057	22809	12	12	12	12	12
■SAL_CD5...			202507895	663989	663989	657950	143479	307	2	2	2	2	2	2
■SAL_CD5...			202507898	663985	663985	663985	147160	147160	56084	27753	27753	12	12	12

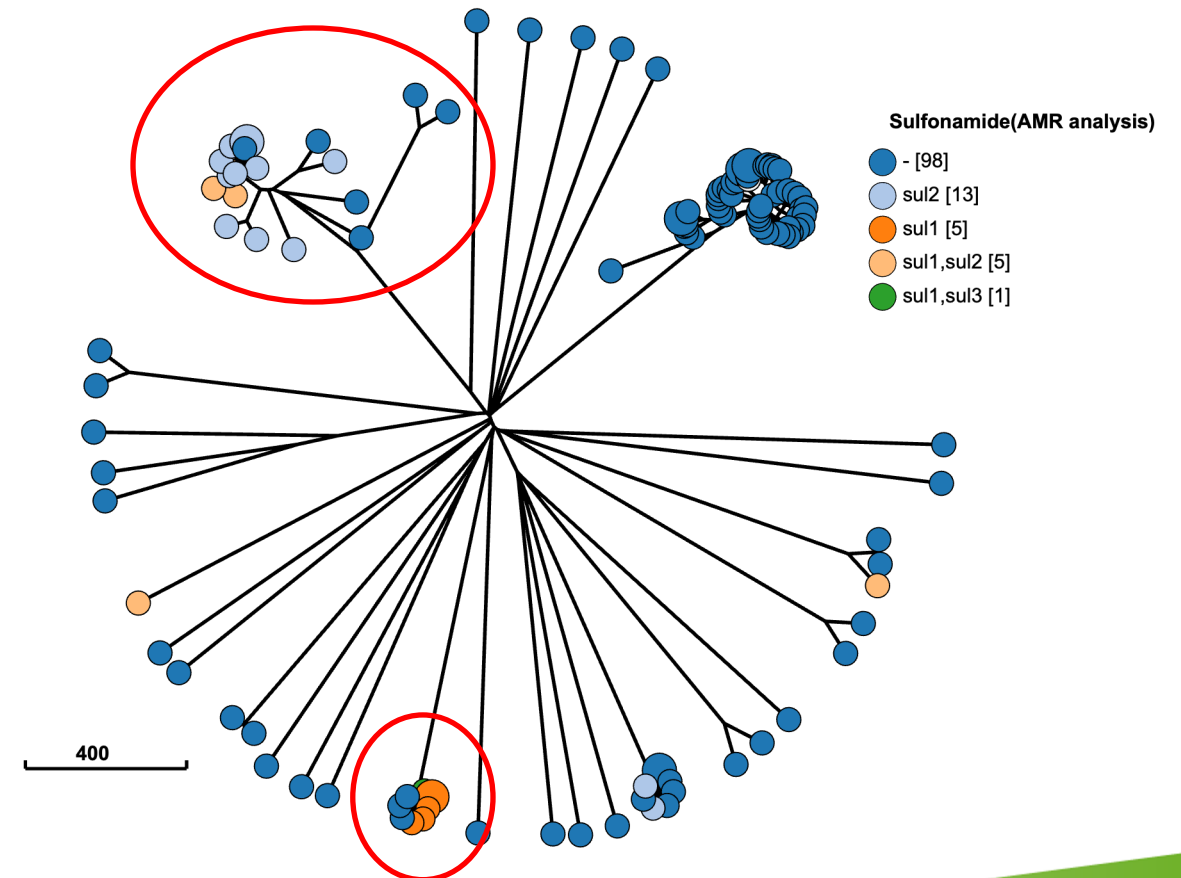
Page size: 50 1-50 51-100 101-123 GoTo: Record Go

Analysis of one day's data from France

HC400 grouping

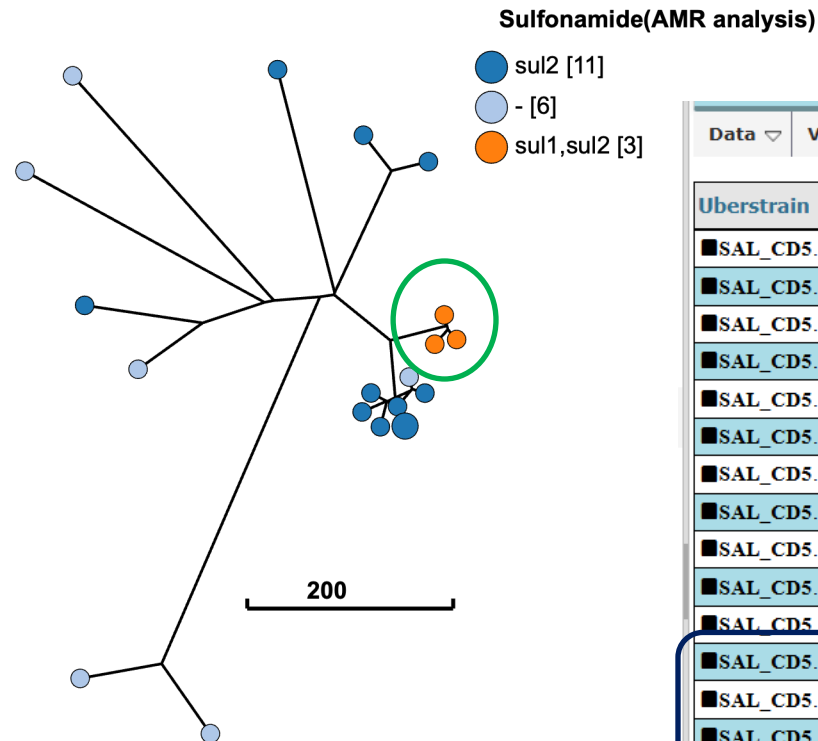


Sulfonamide drug resistance



Zoom in on samples of interest

Appears to contain three samples with multi-drug resistance



Data ▾			View ▾			Workspace ▾			Workspace:None			Rows Total:20 Filtered:20			Edit Mode: ☐ Experimental Data			AMR analysis		
Uberstrain	✓	✎	Name	Penicillin	Phenicol	Quinolone	Sulfonamid...	Tetracycline	Trimethopr...	Other Classes										
■SAL_CD5...			202507895	blaTEM-1	-	-	sul2	tet(B)	-	-										
■SAL_CD5...			202507879	blaTEM-1	floR	qnrB19	sul2	tet(A),tet(B)	dfrA12	Lincosamide:lnu(F)										
■SAL_CD5...			202507948	blaTEM-1	-	-	sul2	tet(B)	-	-										
■SAL_CD5...			202507951	blaTEM-1	-	-	sul2	tet(B)	-	-										
■SAL_CD5...			202507994	-	-	-	sul2	tet(A)	-	-										
■SAL_CD5...			202507970	blaTEM-1	-	-	sul2	-	-	-										
■SAL_CD5...			202507965	blaTEM-1	-	-	sul2	tet(B)	-	-										
■SAL_CD5...			202507982	-	-	-	sul2	tet(A)	-	-										
■SAL_CD5...			202507973	blaTEM-1	-	-	sul2	-	-	-										
■SAL_CD5...			202508027	blaTEM-1	floR	-	sul2	tet(B)	-	-										
■SAL_CD5...			202508023	blaTEM-1	-	-	sul2	tet(B)	-	-										
■SAL_CD5...			202507884	-	floR	qnrB2	sul1,sul2	tet(A)	dfrA1	-										
■SAL_CD5...			202507911	-	floR	qnrB19	sul1,sul2	tet(A)	dfrA1	-										
■SAL_CD5...			202507989	-	floR	qnrB2	sul1,sul2	tet(A)	dfrA1	-										
■SAL_CD5...			202507908	-	-	-	-	-	-	-										
■SAL_CD5...			202507917	-	-	-	-	-	-	-										
■SAL_CD5...			202507995	-	-	-	-	-	-	-										
■SAL_CD5...			202507966	-	-	-	-	-	-	-										
■SAL_CD5...			202507971	-	-	-	-	-	-	-										
■SAL_CD5...			202508002	-	-	-	-	-	-	-										

Click through to
associated NCBI data

GenPept

Send to:

Change

Custom

This record is a non-redundant protein sequence. Please [read more here](#).

MULTISPECIES: sulfonamide-resistant dihydropteroate synthase Sul2 [Pseudomonadota]

NCBI Reference Sequence: WP_001043265.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Download Datasets

Go to: ☐

LOCUS

WP_001043265

271 aa

linear

BCT 03-MAR-2025

DEFINITION

MULTISPECIES: sulfonamide-resistant dihydropteroate synthase Sul2 [Pseudomonadota].

ACCESSION

WP_001043265

VERSION

WP_001043265.1

KEYWORDS

RefSeq.

SOURCE

Pseudomonadota (proteobacteria)

ORGANISM

[Pseudomonadota](#)
Bacteria; Pseudomonadati.

COMMENT

REFSEQ: This record represents a single, non-redundant, protein sequence which may be annotated on many different RefSeq genomes from the same, or different, species.

##Evidence-For-Name-Assignment-START##
Evidence Category :: Antimicrobial Resistance Protein
Evidence Accession :: [NG_048118.1](#)
Evidence Source :: Bacterial Antimicrobial Resistance Reference Gene Database
##Evidence-For-Name-Assignment-END##
COMPLETENESS: full length.

FEATURES

Location/Qualifiers

source

1..271

/organism="Pseudomonadota"

/db_xref="taxon:1224"

gene

1..271

/gene="sul2"

Protein

1..271

/product="sulfonamide-resistant dihydropteroate synthase Sul2"

/calculated_mol_wt=28397

Protein

Sulfonamide-resistant dihydropteroate synthase Sul2

Total protein conservation

Related

[BioProject](#)
[Nucleotide](#)
[Taxonomy](#)
[CDD Search](#)
[Conservation](#)
[Conservation](#)
[Domain](#)
[Genome](#)
[Genomic](#)
[Protein](#)

Identify hc10 cluster and find other entries

Search Strains ¹

Maximize

Data ▾ View ▾ Workspace ▾ Experiment ▾



Workspace:None Rows Total:3 Filtered:3

Edit Mode: ☐

Experimental Data cgMLST V2 + HierCC V1 ▾

Uberstrain	✓	✎	Name	Data Source	ST	HC0 (in...	HC2	HC5	HC10	HC20	HC50	HC100	HC200	HC
■SAL_CD5...			202507884	⊕Uploaded R	663991	663991	663991	649086	1499	260	260	2	2	2
■SAL_CD5...			202507911	⊕Uploaded R	663997	663997	1499	1499	1499					
■SAL_CD5...			202507989	⊕Uploaded R	664105	664105	328054	1499	1499					

Undo Changes in Cell

Select All

Unselect All

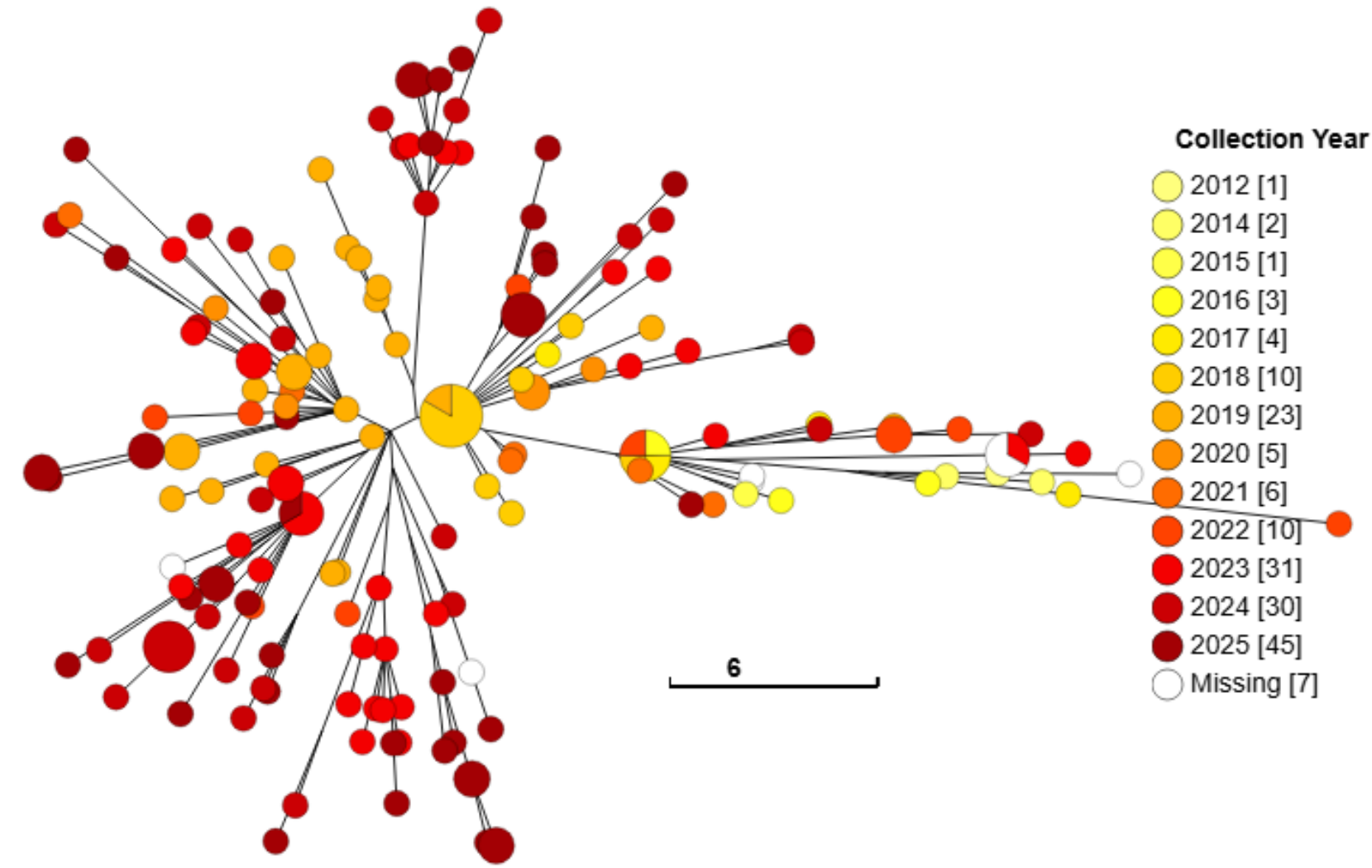
Find ST(s)

Get at this level

Download Allelic Profile

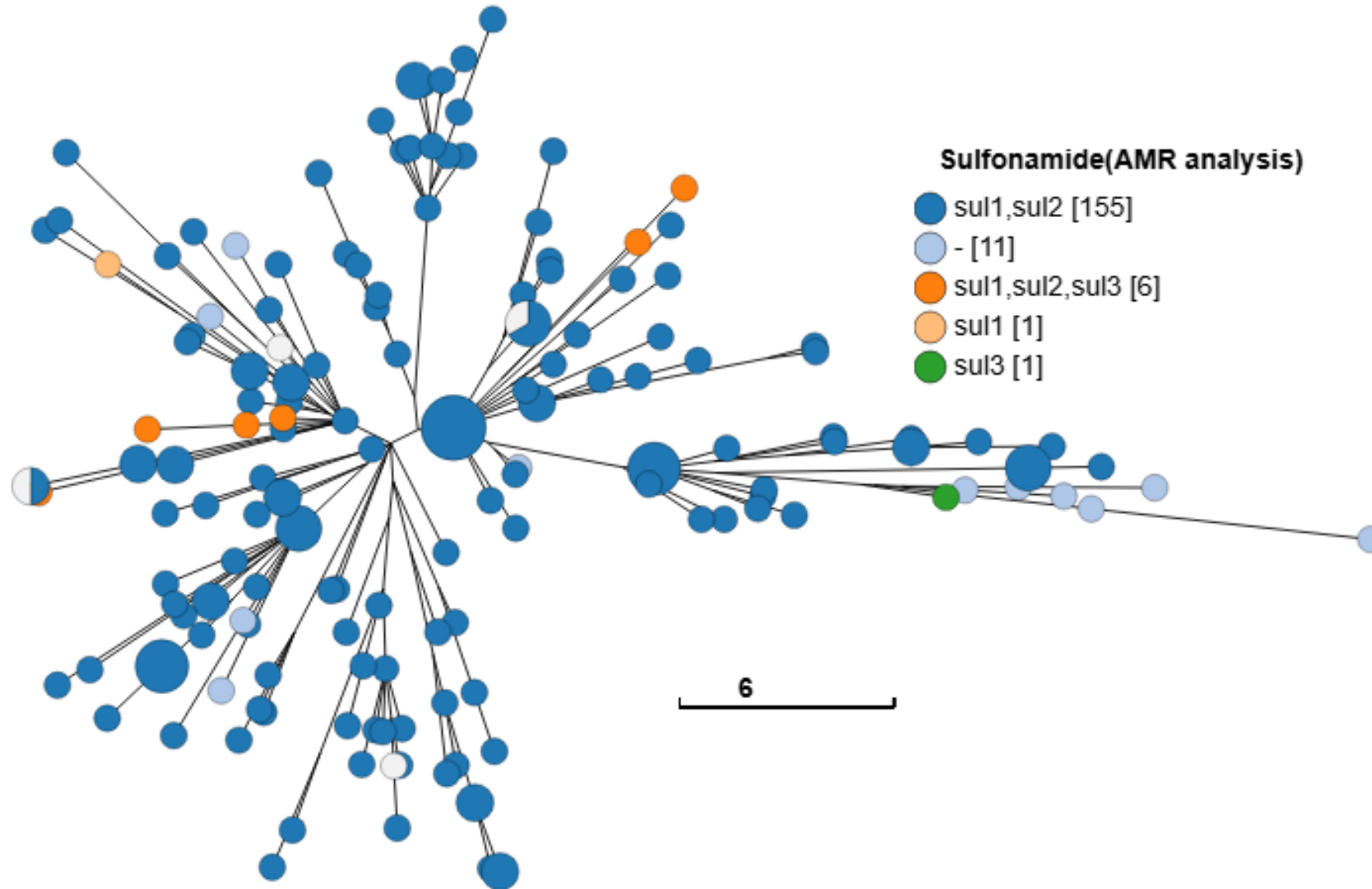
Download Allelic Profile (Selected Only)

hc10_499 goes back to 2012



Enterobase holds
178 samples in this
cluster, going back
to 2012

Long-standing sulfonamide resistance



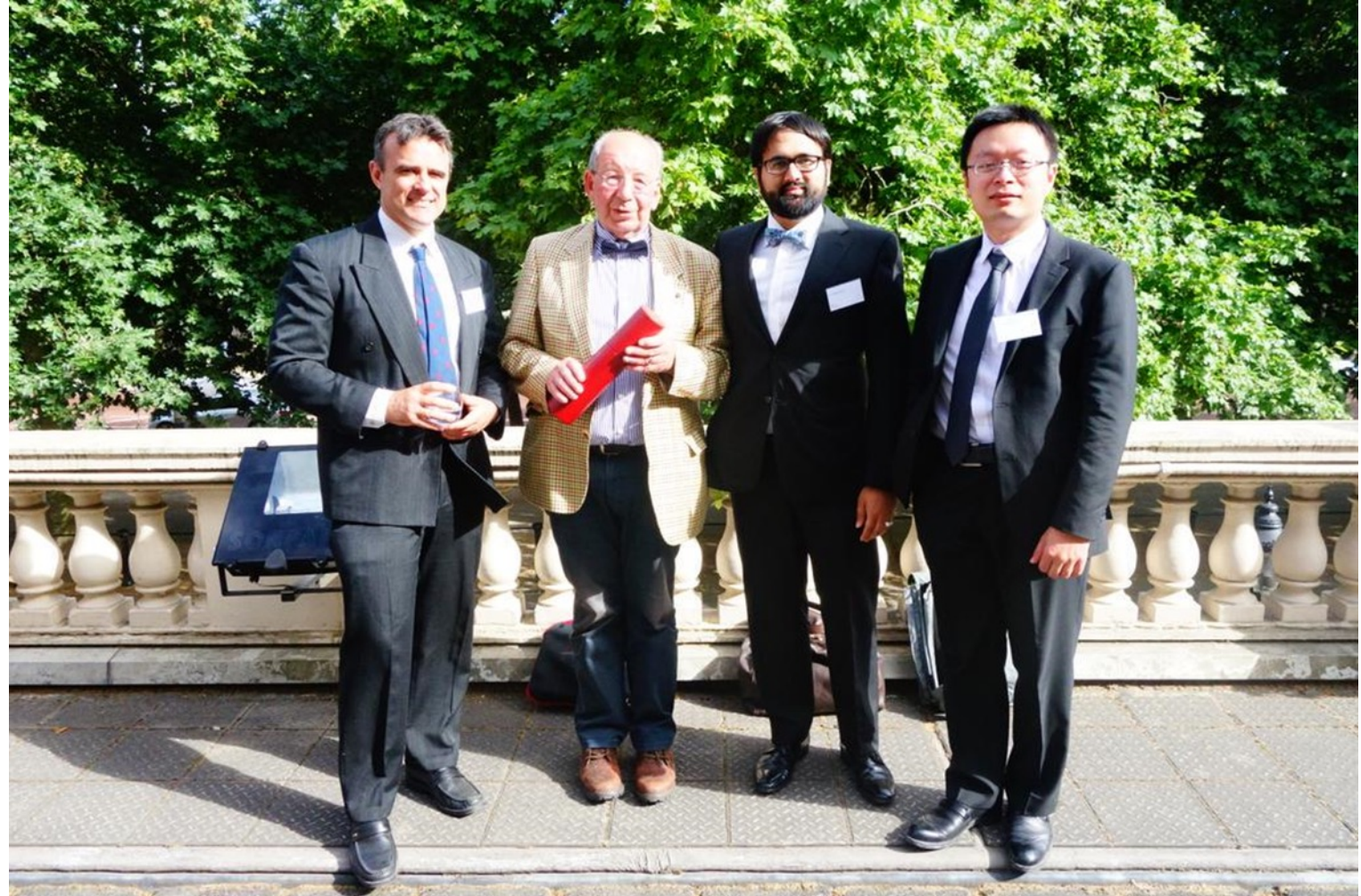
References

- Zhou Z, Alikhan NF, Mohamed K, the Agama Study Group, Achtman M (2020). The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity, [Genome Research, 30:138-152](#)
- Zhou Z, Charlesworth J, Achtman M. (2021). HierCC: A multi-level clustering scheme for population assignments based on core genome MLST, [Bioinformatics, 37\(20\):3645–3646](#)
- Nigel P Dyer, Birgitta Päuer, Laura Baxter, Anshul Gupta, Boyke Bunk, Jörg Overmann, Margo Diricks, Viola Dreyer, Stefan Niemann, Kathryn E Holt, Mohammed Rahman, Paul E Brown, Richard Stark, Zhemin Zhou, Sascha Ott, Ulrich Nübel, EnteroBase in 2025: exploring the genomic epidemiology of bacterial pathogens, *Nucleic Acids Research*, 2024, <https://doi.org/10.1093/nar/gkae902>
- <https://enterobase.readthedocs.io/en/latest/>

Acknowledgements

Original Developers:

- Mark Achtman
- Zhemin Zhou
- Nabil-Fareed Alikhan
- Martin Sergeant



Acknowledgements

Current Team Members



Laura Baxter

Nigel Dyer

Richard Stark

Paul Brown

Sascha Ott



Birgitta Päuer, Anshul Gupta, Boyke
Bunk, Jörg Overmann, Ulrich Nübel

References

- Prokka Annotation: <https://github.com/tseemann/prokka>
- AMRfinderPlus: <https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/>
- ClermonTyping: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6113867/>
- EZClermont <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7656184/>

Acknowledgements

The creation of this training material was commissioned by ECDC to the University of Warwick with the direct involvement of Dr Nigel Dyer